Food | Consumer | Health
Designing a world-class infrastructure to facilitate research

Horizon 2020
INFRADEV-1-2014 - Design studies

RICHFIELDS Working Package 9
Deliverable 9.3

Scientific manuscript on overall case study outcomes and future framework

Date delivered:
M36

Authors:
Mark Roe, Rachel Berry, Barbara Koroušić Seljak, Tome Eftimov, Tamara Bucher,Julie-Anne Nazare, Martine Laville, Angelika Mantur-Vierendeel, Todor Ginchev, Jose Costa-Requena, Sophie Hieke, Heinz Freisling, Paul Finglas

Deliverable lead beneficiaries:
QIB (formerly IFR)

www.richfields.eu
#RICHFIELDS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

Project	
Project acronym:	RICHFIELDS
Project full title:	Research Infrastructure on Consumer Health and Food Intake for E-science with Linked Data Sharing
Grant agreement no.:	654280
Project start date:	01.10.2015
Document:	
Title:	Scientific manuscript on overall case study outcomes and future framework
Deliverable No.:	D 9.3
Authors:	Mark Roe, Rachel Berry, Barbara Koroušić Seljak, Tome Eftimov, Tamara Bucher, Julie-Anne Nazare, Martine Laville, Angelika Mantur-Vierendeel, Todor Ginchev, Jose Costa-Requena, Sophie Hieke, Heinz Freisling and Paul Finglas
Reviewer:	Karin Zimmermann – Project Coordinator prof. dr. ir. Pieter van't Veer – Scientific Coordinator
Start date:	01.10.2015
Delivery date:	29.10.2018
Due date of deliverable:	31.07.2018
Dissemination level:	Public
Status:	Final

Change history:		
Version	Notes	Date

Karin Zimmermann

Project Coordinator

Prof. Pieter van't Veer

Scientific Coordinator

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

Summary

The overall goal of the RICHFIELDS Horizon 2020 project is to bring together the agri-food and nutrition-health domains to collect, collate and connect relevant data, tools and resources to create a single, multidisciplinary Research Infrastructure (RICHFIELDS Data Platform). In order to create a relevant, efficient, effective and fit-for-purpose design, current and emerging RIs and related entities (tools, networks, platforms) were evaluated to assess the feasibility of implementing the RICHFIELDS Data Platform. The evaluations were conducted as four case studies focussed around three related research areas: determinants of dietary behaviour, intake of foods and nutrients, and status and functional markers of health. The assessments used common questions of evaluation, covering data structure, ethics, user needs and business models. In a related exercise, several tools that are currently under development, but that fit the core offering of the RI, were also tested to further determine potential integration with any future platform. The aim of the case studies and tool evaluation was to make recommendations for the future development of a RICHFIELDS Data Platform, and to highlight opportunities and challenges in building the RI. The case studies provided both specific and over-arching conclusions and recommendations across the evaluation themes. Although the assessment themes are reported individually, there is considerable influence and overlap between them (e.g. ethical requirements drive data handling procedures), which will ultimately help to shape the design and functionality of the RICHFIELDS Data Platform. New tools and resources that are relevant to the RI are continually being developed, and the platform would benefit from their inclusion or connection, therefore it is important to engage with developers at an early stage in order to promote the sustainability and maintain relevance of the RI.

.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

1. Introduction

This deliverable is a draft outline of a manuscript for submission for publication in a peer-reviewed scientific journal in a field that is relevant to RICHFIELDS, such as the Public Health Nutrition journal. Alternatively, the European Journal of Clinical Nutrition or Nutrients may be interested in the topics of the publication, or run specific issues on related topics. A final option would be to try to publish in a journal more specific to food composition, such as the Journal of Food Composition and Analysis. The manuscript is based on the case studies and outcomes reported in deliverables 9.1 and 9.2 and outlines a possible future framework for handling data, made available by data providers and shared with users, that could exist within the RICHFIELDS Data Platform RI.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

2 Background

The overall aim of the EU funded 'Research Infrastructure on Consumer Health and Food Intake using E-science with Linked Data Sharing' (The RICHFIELDS Project, 2018) project is to design a world-class infrastructure (RICHFIELDS Data Platform) for innovative research on food preparation and consumption, and healthy food choice, in EU citizens, closely linked to their behaviour and lifestyle. Data related to food choice, nutrition and health is fragmented, key information is lacking, and the resulting knowledge gap limits policy makers in making effective public health nutrition strategies, and companies in exploiting current scientific evidence. RICHFIELDS is intended to bridge the gap by linking the agri-food and nutrition-health domains and account for the regional and socio-economic diversity of the EU. The infrastructure will collate, connect and collect data, tools and resources to enable multidisciplinary research, in both the public and private sectors, to support the development, implementation and evaluation of effective food and health strategies.

High quality data that is easily available and that can be adapted for a wide range of uses by users with different backgrounds and levels of expertise is essential to support a successful research infrastructure. Tools to support both providers of data and users of data by enabling efficient exchange and transformation of data are also essential. The RICHFIELDS project identified that the main groups of data users and providers are: researchers/academia; public health policymakers; healthcare professionals; commercial (agri-food industry, retail, software); and consumers.

The RICHFIELDS project identified, analysed and tested the feasibility of implementing the RICHFIELDS Data Platform infrastructure, particularly focusing on linking data, resources and supporting information, both in terms of the platform content and the technical aspects. Current and new Research Infrastructures (RIs) in the field of food and health were examined to consider how they may be linked to a RICHFIELDS Data Platform, considering issues such as data structure, ethics, user needs and business models. Research Infrastructures can be defined at different levels, with very few meeting the strictest definition set out by the European Strategy Forum on Research Infrastructures (ESFRI) and attaining European Research Infrastructure Consortium (ERIC) status (Brown et al 2017). In selecting the case-studies for this task, guidance was taken from the entities identified during the EuroDISH project using a softer definition of a RI (European Commission, 2018), and the platforms, tools or structures that may be incorporated under the umbrella of the RICHFIELDS Data Platform and directly share or provide data. The aim of this work was to identify opportunities and challenges and make recommendations for development of the platform to enable the ability to include or link to data that is either generated within RICHFIELDS or provided to RICHFIELDS by data producers.

3 Methods

Four case-studies were identified at the onset of the project, representing cases related to determinants of dietary behaviour, intake of foods and nutrients, and status and functional markers of nutritional health. The case studies were designed to review the current position and gaps and needs in relation to food composition data, dietary intake for population-based intake assessments, clinical intervention studies, and diet, health and lifestyle information. The case studies selected represent types of data that could be provided by external data providers, or that could be generated directly by users of the RICHFIELDS platform. These types of data were also considered in view of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

likely use and value for the key target user groups of the RICHFIELDS Data Platform. The outcomes of the EuroDISH project, which identified RIs and related entities (e.g. platforms, networks and tools) relating to Determinants, Intake, Status and Health (DISH) research areas (Brown et al 2017), were used to guide the selection of the case-studies, along with the expertise of the work-package partners.

The main phase of each case study consisted of evaluation of relevant RIs, datasets, and tools to answer questions that are essential for the development of a consumer data platform:

- **Data structure:** How are the selected food composition/food consumption/clinical or lifestyle datasets used within relevant applications structured?
- **Data storage and availability:** How are information technologies used to make the data available to users and how and where is the data stored?
- **Maintenance and access:** How do the data producers/compilers who maintain these datasets evaluate data access, exchange and linkage to external RIs? What would be the challenges and constraints to expand access to the data?
- **Ethical issues:** What are potential ethical issues related to linking into a RI (e.g. data privacy, ownership rights etc.)?
- **Recommendations for RI design:** What recommendations can be made on the design of future data structures and interfaces of datasets and applications, taking into account a pan-European research infrastructure as proposed by RICHFIELDS?

In addition, several tools that could be relevant for the RICHFIELDS Data Platform were tested and evaluated as part of their ongoing development. This activity was designed to highlight potential design and performance considerations (particularly relating to data) of the relevant tools if linked with the RICHFIELDS Data Platform, both from the tool developers' prospective, and that of the RI.

3.1 Case study 1: Food Composition and food attributes

The aim of the case study was to explore and develop techniques and tools to support the linkage and the harmonization of diverse food composition data coming into the future RICHFIELDS Data Platform from different datasets. A variety of tools related to use of food composition data have been developed in recent years, and the case study particularly focussed on tools that enable access to harmonized and standardized data and tools that enable food composition data to be combined with other data types, e.g. food consumption data. Food composition datasets, including national datasets and specialized datasets for bioactive compounds and branded foods were assessed in terms of their availability, degree of harmonized data structure, limitations of use and potential for linking to the RICHFIELDS Data Platform.

Tools that are under development were also evaluated for their potential use in the RICHFIELDS Data Platform, including tools for food matching (to enable food data classified and described using different systems to be linked) and tools for identifying and estimating the weight of foods consumed (Table 1). Specific tools included: pocket size Bluetooth weighing scales (Papa et al 2015); use of Semantic Web metadata, using the Quisper ontology (Eftimov 2018), to test data identification and extraction of data using the Part-of-Speech tagging-probability weighted method (Eftimov and Koroušić Seljak, 2015) and to test automatic annotation of data using the StandFood method (Eftimov et al., 2017); food portion image detection and recognition using the NutriNet system (Mezgec and

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

Koroušić Seljak, 2017); and the Fake Food Buffet tool to measure food product and portion size choice in laboratory studies (Bucher et al 2012). These tools were tested based on their current level of development and the potential for further development and inclusion to support the RICHFIELDS Data Platform.

3.2 Case study 2: Standardized food intake from population-based survey

The type of data related to food intake and consumer behaviour that could be collected in the frame of dietary monitoring systems conducted across Europe using a standardized dietary methodology was identified. The concepts of comparable food intake data and relevance for consumer behaviour were tested using a standardized dietary methodology. The GloboDiet platform (formerly known as "EPIC-Soft") was chosen as an example of a well-developed, standardized tool for capturing food consumption data because it has already been adapted for national surveillance systems in several European countries (de Boer 2011), as an integrated part of the EU-Menu project (EFSA 2009, Slimani et al 2011). Existing gaps and needs in the processes used within GloboDiet, which would also be applicable to similar software systems used to capture food consumption, were identified with the aim of improving data linkage among pre-existing and new tools for the RICHFIELDS platform.

3.3 Case study 3: Clinical interventions

The aim of the case study was to identify and map the types of data related to food intake and consumer behaviour that could be collected in the frame of clinical interventions considering the type and design of studies, type of patients and populations, and clinical outcomes. A range of relevant clinical intervention studies were reviewed and networks were identified that were used to provide information on the types of studies and data that may be available to link to RICHFIELDS. Gaps and needs for linking clinical data with the RICHFIELDS Data Platform, particularly data management and legal and ethical constraints, were evaluated to assess the feasibility of data exchange from clinical data sources.

To further discuss current practices in data collection and data sharing, a meeting was organized in Paris in June 2017 between representatives of identified RIs of interest to share and discuss RIs' structures and developments related to clinical data collection, integration, sharing, access and related technical ethical, legal issues in the nutrition area. RIs involved in the meeting included RICHFIELDS, ECRIN, ECRIN Nutrition, ENPADASI and CORBEL/ELIXIR. The aims of the meeting were to: discuss and identify areas of possible collaboration between RICHFIELDS, and other RIs; share best practice and approaches in existing RIs that would be useful for RICHFIELDS and; agree plans and content to provide information for the case study.

3.4 Case study 4: Consumer diet, health and lifestyle

Delivery of data related to dietary intake, health and lifestyle from user apps is a key aim of the RICHFIELDS Data Platform. This case study explored the potential for delivering data and content from various features of the PRECIOUS RI (The Precious Project, 2018) and is a 'proof of principle' of the features and flow of data that could be used to support other similar applications. The PRECIOUS platform (Mutafungwa et al 2015) produced a smartphone application that collects information about the user from a variety of devices and applications (sensors) that measure food intake, physical activity, stress levels and sleep patterns. These tools were used to evaluate usability, efficiency and

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

acceptability by groups of users and software application developers, and to assess what further developments would be needed for implementation in the RICHFIELDS platform. The features considered included backend architecture design and use of a backend API for an application to collect data. Usability of the PRECIOUS platform was evaluated from three different user viewpoints: user-related studies of usability, efficiency and acceptance; sensing and home-environment related studies of how the platform could contribute to health improvements of users; studies assessing the possibility to include third party developers to add to functionality of the platform.

3.5 Identification of needs for RI development

Each case study was reviewed and issues that would be important for development of a RI were identified. Some issues were case study specific and others were identified as overarching issues that would be central to development of effective strategies for linking to and sharing data between data providers and users. These issues were the basis of recommendations, based on the case study outcomes, for development of the RICHFIELDS Data Platform. The recommendations were provided to the RICHFIELDS project management team for consideration in the wider context of data integration and data management to enable creation of a functional open-access, distributed platform to empower exploration and exploitation of food-related data collected from both European consumers and existing RIs via different data sources. Some of the issues explored were also related to design of a sustainable business model for the Platform that provides value to stakeholders. These business sustainability related issues were considered and further developed by the RICHFIELDS project.

4 Results

4.1 Case study 1: Food Composition and food attributes

4.1.1 Food composition data

Evaluation of available food composition datasets showed that most national food composition datasets are freely available to use, even for commercial purposes, with only a few that are subject to license agreements. National nutrient datasets are usually managed and stored centrally but, in many cases, are available for searching or download from the national dataset curators and/or publishers. Data is usually available in standard formats (e.g. spreadsheet, database, xml) that could be incorporated into a data platform. Documentation to describe datasets, including information on data sources and definition of components is available for most datasets and can help inform standardization and comparison of data.

EuroFIR (EuroFIR, 2018) and INFOODS (INFOODS, 2018) have both developed standards for production of harmonized nutrient data, and national datasets are becoming more harmonized as more national compilers are trained in the use of standards and are increasingly aware of the need to produce data that is comparable with other datasets. However, accessing and combining data from different datasets is not an easy task, even though the data for individual datasets is easily accessible. The EuroFIR FoodEXplorer tool is an innovative interface, which can be accessed online and allows users to simultaneously search standardized and specialized food composition databases (Finglas et al 2014). The values available within FoodEXplorer are provided by national compilers but are presented

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

in a harmonized format and search results can be downloaded into formats that enable further use. Even though the presentation of results is harmonized, there are still some differences in how values are compiled at national level and further work to fully standardize data for inclusion in a data platform would be needed. EuroFIR has also developed a range of generic tools to make data available to users and to allow transfer of data between users (e.g. Food Data Transport Package, Food Data Query Language (used to request data via webservices). Training data compilers and users to make them aware of data standards and the correct use of tools has also been an important feature of EuroFIR. The EuroFIR membership structure provides access to the network data, tools and expertise and is available to users on an individual or organizational basis.

Nutrient composition of branded food products is widely available but most sources do not allow easy assimilation into other datasets. There are commercial sources of some retailer and manufacturer datasets but they are expensive to license. Branded food data is accessible through retailer and manufacturer websites, usually for individual food products. Some data sources (e.g. Tesco (UK), Brandbank) have made APIs available to users to allow product information, including nutrient content, to be embedded into software applications. The CEN standard for food data (European Committee for Standardisation, 2012) is compatible with the GS1 system for barcodes and therefore it is possible for branded food data to be linked to both food composition and food consumption data.

Data on bioactive components in food is also available through EuroFIR via the eBASIS database (Plumb et al 2017). Data on polyphenols can also be found in the Phenol-Explorer database, which is an online database with free and unrestricted access for all users. The USDA bioactive databases are also freely available.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

4.1.2 Tools for linking food composition and food intake data

Food matching tools and food classification and description systems for linking composition data to other food information data (including intake, purchasing and preparation) were evaluated. These systems are essential tools for linking between datasets and platforms and for enabling high quality data searching and data retrieval. As the future RICHFIELDS Data Platform will need to link diverse data coming from diverse datasets, food matching tools and food classification and description systems for linking, for example, food composition data to consumption data and potentially other determinants data are essential. Evaluation of the tools assessed in case study 1 showed some potential for inclusion of such tools within the RICHFIELDS platform.

The Jozef Stefan Institute, Slovenia, (JSI) designed and developed a weighing scale, named Libra, as a simple and inexpensive measuring device for real time portion size assessment (Papa et al, in press). Libra is based on a generally available kitchen scale, where a specially designed communication module is added to support Bluetooth communication with a mobile app (Nutri – developed by JSI). The Libra Bluetooth module is connected to the output of the scale's electronics, so that the existing electronic parts can operate without any changes. The upgraded kitchen scale is a device with a new electronic circuit for the communication module, which can be reused for any generally available kitchen scale that does not have a built-in communication module. The prototype scale was further developed during the RICHFIELDS project to achieve a smartphone size and appearance, while still able to fit all the mechanical (i.e., weight sensors, capacitance button) and electronic (i.e., scale controller, communication module, battery) parts inside. The usability of the scale was tested by five volunteers, diabetes patients from a Slovenian general hospital, who evaluated the scale and the app at home. After using the scale and the app to record food and drinks for 21 days, the patients completed a self-administered questionnaire. The results of the feasibility study proved the simplicity and efficiency of the Libra scale and the mobile app Nutri. Furthermore, the linkage of the Libra scale to the FoodWiz2 app (designed to capture information on food intake and physical activity) was tested by QIB as part of the QuaLiFY (Quantify Life Feed Yourself) project. The app and scale were used to capture portion sizes by a group of 34 adolescents in the UK who took part in a study to test usability of FoodWiz2 (Jimoh et al., 2018). In addition, an upgraded version of the scale was produced to improve efficiency and Android and iOS apps were developed to perform weight measurement and Over-The-Air software update by smart phones.

JSI tested linking data from GS1, PRECIOUS and EuroFIR to the RICHFIELDS ontology using the StandFood method (Eftimov et al., 2017b). The GS1, PRECIOUS, and EuroFIR data is represented as semi-structured data (e.g., xml, json) and the aim was to automatically annotate the food concepts presented in these documents with the links of the concepts that exist in the RICHFIELDS ontology. Two scenarios were performed. In the first scenario, by linguistic pre-processing of the semi-structured data, the food concepts were selected and annotated using the RICHFIELDS ontology. The second scenario was more challenging because the text phrases from the documents that represent food concepts need to be automatically recognized and then annotated. For this reason, the JSI team developed a method known as drNER (Eftimov et al, 2016; Eftimov et al, 2017a), which was developed to automatically recognize food, nutrient, and quantity/unit concepts from unstructured text.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

For each food concept from the GS1, PRECIOUS, and EuroFIR data, StandFood gives a matching probability and the most relevant concept that exists in the RICHFIELDS ontology. If the matching probability is greater than a set probability threshold, then the food concept is annotated with the link from the same concept that exists in the ontology, otherwise the concept does not exist in the ontology and the ontology needs to be populated with the new food concept. The GS1, PRECIOUS, and EuroFIR documents can be changed and annotated using the semantic metadata from the ontology, but also the RICHFIELDS ontology is automatically changed when a new food item that does not exist is added, however the ontology population also needs to be checked by a human expert. This work demonstrated the possibility to work with datasets produced using different standards and represent them in a standard ontology. The RICHFIELDS ontology is an important cornerstone of the future RICHFIELDS Data Platform. It enables not only the harmonisation of data linked by RICHFIELDS, but also its connectivity with other information systems like CORBEL, BBMRI, EXCEL etc. The RICHFIELDS ontology and the methodology for automatic creation of such an ontology can be offered as a product/service as well.

Another approach developed by JSI was based on deep machine learning. To detect and recognize food and drink images, a novel deep learning architecture, named NutriNet (Mezgec and Koroušić Seljak, 2017) was defined. This architecture was tuned on a dataset containing 129,141 images of 520 different food and drink items, on which a 82.11% classification accuracy of 82.311% and detection accuracy of 94.51% was achieved. A real-world test of the recognition model was performed, on a dataset of self-acquired images (using the Google Custom Search API) combined with images from users of a mobile app developed for this aim, and achieved a top-5 accuracy of 52%. A server-side training component was implemented to continually fine-tune the food and drink recognition model on new images. This approach shows promise for accurate recognition of food and drink images but must be further developed to tackle the problem of automatic recognition of the volume of food captured by a low-cost mobile device in a real-life setting.

The fake food buffet (FFB) method is an established and validated tool to measure food product and portion size choice in laboratory studies. The FFB method uses authentic food replicas linked to a nutrient database to measure choice and knowledge. Food replicas are weighted manually and images of selected foods and meals are captured. The use of FFB data to validate data links between composition data and consumer behaviour data (e.g. consumption and preparation) was evaluated by JSI and ETHZ (Zurich, CH). The FFB was tested in combination with a new food matching technology to automate data collection and analysis. The methodology combined fake food image recognition by using Deep Learning and food matching and standardization based on natural language processing. Food matching firstly describes each of the recognized food items in the image and then matches the food items with their compositional data considering both their food names and descriptors. The final accuracy of the deep learning model trained on fake food images acquired by 124 study participants and providing 55 food classes was 89.96%, while food matching performed with a classification accuracy of 93%. More details are provided in a scientific paper prepared by JSI and ETHZ (Mezgec et al, 2018). The study demonstrates that the food matching technology might be useful to automate and standardize food selection data and integrate existing FFB data (captured on food images) into a potential RICHFIELDS research infrastructure.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

4.2 Case study 2: Standardized food intake from population-based survey

GloboDiet is a computerised program to conduct 24-hour dietary recall interviews, either face-to-face or by telephone, and a data entry application for food dairies (Freisling et al. 2015 PMID: 24916012), which has been developed at the International Agency for Research on Cancer (IARC), together with multiple end-users and multidisciplinary partners (Slimani et al., 2000, 2011 PMID: 24916012, Freisling et al. 2015). It is a standardised methodology which has been designed, validated, and implemented as a reference methodology for both European nutritional epidemiology studies and for future Europe-wide nutritional surveillance (de Boer et al., 2011; Slimani et al., 2002, 2011). Review of GloboDiet data structures showed that the software is generally compatible with food composition data and datasets from different countries can be incorporated into the system. The GloboDiet methodology has been successfully implemented in European nutritional epidemiology projects, such as the *European Prospective Investigation into Cancer and Nutrition* (EPIC) calibration study and the *European Food Consumption Validation* (EFCOVAL). In addition, it is used as a common standardised dietary methodology in seven National Surveillance systems in Europe (Austria, Belgium, France, Germany, Malta, The Netherlands, and Switzerland) under the double EU-Menu and GloboDiet umbrella. GloboDiet collects food consumption data that is sufficiently detailed to provide comparable food consumption data (Crispim et al. 2011 PMID: 21731004), nutrient intake (Freisling et al. 2010 PMID: 20484545) and exposure to dietary contaminants (Crispim et al., Freisling et al. 2013 PMID: 23238529). GloboDiet also collects data to assess eating habits or patterns and meal patterns in multi-national settings (Park et al. 2018 PMID: 28275868, Huseinovic et al. 2016 PMID: 27194183).

4.3 Case study 3: Clinical interventions

A range of relevant clinical intervention studies were reviewed and several trials were identified that could be used for the case study to provide information on the types of studies and data that may be available to link to RICHFIELDS.

A completed, FP7-funded, Randomised Controlled Trial from the Centre de Recherche en Nutrition Humaine Rhône-Alpes (CRNH), Lyon, France (clinical partner) for which CENS had full access to data (EUROSTARCH, 2018) was chosen as a good example to study. EUROSTARCH addressed a large panel of nutrition and health parameters and, as a typical complex intervention and phenotyping study, was also chosen as a case-study to evaluate potential integration within both the national FORCE (French Obesity Research Centre of Excellence, 2018) database (meta-data) and the phenotype Database DbNP within the European Nutritional Phenotype Assessment and Data Sharing Initiative (ENPADASI) project (dbNP 2018, ENDAPASI 2018). In addition to studying data structure and data access it is also relevant as a study of how data can bridge research infrastructures, based on work undertaken within ENPADASI.

Improvements and novel resources in phenotyping and technologies (such as “omics”) and the direction of clinical research towards personalized medicine have considerably increased the profiling capacities, the volume and typologies of data. Data encompass raw data, metadata (that describe the other data; that is type, design, duration, location of a study, informed consent, type and methodologies of analysis, of the population, etc.) and aggregated data (preliminary compiles/analysed). Metadata are generally less sensitive in terms of data sharing as long as they

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

remain anonymous and cannot be linked with personal data. Raw data, personal and clinical data, are controlled by various data protection rules and ethical constraints including informed consent and anonymization. Aggregated and analysed data from studies are usually available from scientific literature but the raw data, even if anonymized, is rarely available. Publication of study results is usually a condition of ethical approval for such studies. In response to external requests for secondary data analysis (meta-analysis, review), anonymized data have been transferred and shared to external consortia, however the governance rules and protocols associated with making anonymized data available to researchers are complicated and should ideally be considered during the initial set-up of projects. In many cases, to allow data sharing it will be essential to describe the governance of data and any possible follow up use as part of the research ethics application for the project.

The EUROSTARCH project intervention treatments and measurements/examinations that are done throughout the study are recorded in a study design calendar and treatments performed are described for each event, including information on the types of samples collected. Data were not collected through applications but directly through self-reported questionnaires, assays from collected biological samples and examinations. Recording of each measurement is very important because it determines the number of samples (and therefore results) that will be uploaded to the database. Most clinical studies would be expected to record subjects, samples taken and measurements made on them in a similar way. Database format should not be an important issue because any database (or spreadsheet) system used should be compatible with easy data manipulation and import/export to other formats, e.g. xls, mdb, sql, xml, cvs.

Evaluation of selected networks that collect clinical research data showed that data harmonization is very limited and usually only within projects. Table 2 provides an overview of the focus and data outputs of RIs and networks related to data standardization and sharing. There are no formalized guidelines for retrospective data harmonization to enable re-use of data and sharing of clinical-related data is intrinsically linked to ethical, social and legal considerations and requirements, due to the sensitivity of the data. Making clinical data available and usable for novel purposes by a wide community of researchers and beyond has become a critical challenge. National and international policies, privacy, practical and regulatory issues are necessarily linked with clinical data collection and use, and use of human data is particularly sensitive. Privacy and confidentiality are major issues and identifying biomedical data is particularly risky in terms of potential stigmatization of individuals. Anonymization and harmonization of data both require time, technical and scientific investments. Thus, many protections should be set-up, that may ultimately complicate researchers' and clinicians' work and restrict data access. In this sense, we also know that many null or negative results and related data are generally not available because they are not published. Lately, several online tools have been developed to provide support about the European legal environment and requirements for sharing clinical data (for example: BioMedBridges Legal Assessment Tool (LAT) (BioMedBridges 2018), BBMRI legal WIKI). These tools gather regulatory, ethics information, texts and documentation. BBMRI has also produced a Policy for Access to and Sharing of Biological Samples and Data (BBMRI-ERIC 2018), that could inform the requirements for a similar policy for RICHFIELDS.

The on-going works from ENPADASI WPs, particularly on structured and standardized data storage, actions required for clinical data sharing, could be the basis for future RICHFIELDS development for

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

clinical data integration. The ENPADASI project has specifically focused on the general rules to share and reuse data based on EU national policies, addressing ethics, data protection, data sharing policies and intellectual property. Within the ENPADASI project, the design of dedicated data sharing infrastructure was planned with related analysis of the data sharing issues. RICHFIELDS will benefit from all this work that has been specifically dedicated to clinical data and bridges that could be built between RIs.

4.4 Case study 4: Consumer diet, health and lifestyle

The PRECIOUS system collects information about the user from a variety of devices and applications (sensors) that measure food intake, physical activity, stress levels and sleep patterns. Information is collected using the platform designed and implemented in the PRECIOUS project and consists of a set of sensors and a mobile application that seamlessly collects all the information from consumers to be processed to determine the user lifestyle. This information is used to identify consumer behaviour and the process takes into account the type of users and populations by RICHFIELDS and clinical outcomes would be in line with the RICHFIELDS Data Platform.

Smartphone technology enables inclusion of low-power sensors (i.e. accelerometers) which, with very little processing power, are capable of recognizing user physical activity and detecting walking, running, standing still, cycling or even travelling in a vehicle. This enables effortless tracking of the user without the need of devices dedicated to that purpose. Sub-applications have been developed to track that activity and collect and present data in an interactive way that with the aim of positively modifying user behaviour. The PRECIOUS smartphone application gives users access to record physical activity, diet and motivations for a healthier lifestyle. Food intake is captured by a 'My Food Diary' sub-application that uses country-specific food items that are collected from national food composition databases that have been curated and properly verified by EuroFIR. When more accuracy or features are needed, there is a possibility to expand the platform functionality by using environmental sensors, wearable sensors such as smart wristbands or high accuracy electrocardiogram (ECG) sensors. The smartphone application is composed of sub-applications that offer easy and fast access to different components that target mainly physical activity, diet and motivational interviewing. The platform comprises a tool for image recognition for food analysis integrated in a mobile application with the data processing in the backend of the application. The food information and user's activities are analysed in backend servers to deliver data semantics models or Virtual Individual Models to produce recommendations for food intake and exercise.

One of the novelties in the PRECIOUS server architecture is the Food Recognition Server. It is an implementation of a machine learning algorithm, carried out by Aalto (FI), for an image recognition and localization of a dish in an image (Figure 1). The model is trained by the UEC 256 food dataset and is able to recognize 256 different classes of dishes. The existing network architecture was modified in order to classify 256 food categories and predict the food item, with the network further modified and trained using fine-tuning approach. The faster algorithm returns a bounding box and confidence score of each of the detected items. Based on the confidence score returned by the algorithm the misclassified results are analysed for the post-processing of the results. Using recommendations with smaller confidence scores improves the accuracy of food recognition.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

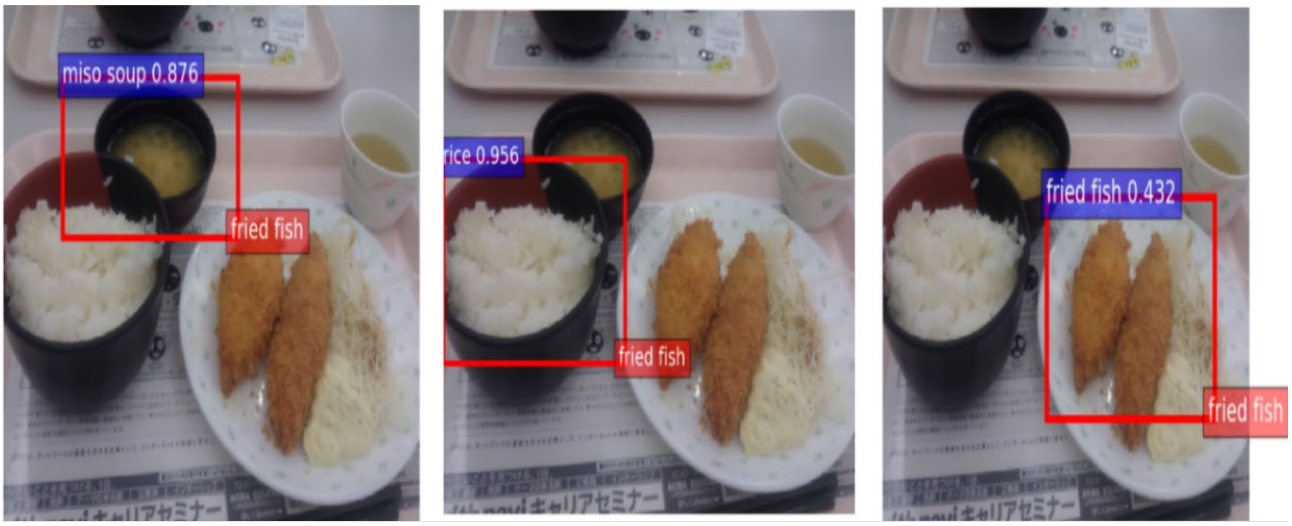

Figure 1. Example of meal image detection and recognition

During the Precious project, Aalto implemented a backend that served the mobile client by authenticating users, securely collecting biometric data and storing it for further processing. While the architecture used was suitable for field trials, it is not suitable for commercial applications. There was therefore a need to redesign the backend architecture to make it more flexible, dynamic and suitable for commercial use. As a starting point, it was decided to collect different kinds of biometric data (e.g. nutrition, physical activity, sleep, etc.) in separated and dedicated micro-servers. This not only enables clear biometric data filtering but it also allows enabling/disabling of services (when there is a need to collect only data related to stress, for example) or the possibility to have part of the backend in a different physical location.

Evaluations identified that smartphone tools are potentially useful for improving health related behaviours but the main criticisms were related to usability aspects. For software developers, there was a need for tools for rapid prototyping and tools for data abstraction and interpretation to aid development of tools for assessing diet, health and lifestyle. Collecting and presenting data is relatively straight forward from a technical point of view but using the data to produce meaningful results can be challenging because of the complexities of data and data interpretation. The Precious System is a potential proof-of-concept for collecting and using real-time data on individual diet and behaviour using a range of different apps, wearables, and sensors. Applications related to the PRECIOUS platform are freely available. The PRECIOUS app has been recognized and registered with the TicSalut Foundation health app observatory in Spain. The app can be downloaded from the google app store (The PRECIOUS Project, 2017).

The case study demonstrated the potential for tools that capture information from users, such as an image recognition tool, to be included in a mobile application aimed at consumers. The link between data collected (front end) and data handling tools, e.g. algorithms to produce dietary or other lifestyle recommendations, demonstrates that there is an option for sophisticated data handling tools to be included in applications that can be used by consumers or researchers to provide data to the RICHFIELDS platform.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

5 Discussion

Results from the four case studies were evaluated and used to identify issues related to the key questions that are essential for the design and development of the RICHFIELDS Data Platform.

5.1 Data structure

There have always been challenges in harmonizing approaches to acquisition and publication of nutrition, health and lifestyle data and while data processing and publication in electronic form makes large datasets easier to handle and more accessible, the challenges related to data quality, data exchange formats and documentation have increased. Many projects and research networks have initiated standardization of methods to collect, manage and publish data but data standardization has not kept pace with fast moving developments in information and communications technology. The ability to generate and exchange large amounts of data has therefore highlighted existing limitations in data structures. A standard classification structure is an essential starting point that can enable datasets to be exchanged and compared. It is not essential for the same taxonomies to be used (although it does make data handling easier) and indeed selection of an appropriate taxonomy, usually hierarchical, may depend on the application. Classifications for food identification and description are used in food composition and food consumption data and various taxonomies are available for use with data related to health (e.g. Medical Subject Headings (MeSH)), and lifestyle (e.g. Compendium of Physical Activities, Ainsworth et al, 1993).

Traditionally there has not been a standard structure for food composition data because datasets have been compiled independently for publication in country specific printed tables in books and scientific journals. Since the introduction of computerized data compilation and publication, there has been a trend towards more standardized data structures and control of data quality through clear documentation of the data. Food composition data can be derived from a range of sources and most food composition datasets contain values produced in a variety of different ways. Nationally representative datasets are produced in most developed countries and increasingly in developing countries, although coverage of foods and nutrients may be more limited. The quality of published data is high and the data are typically produced, managed and published by groups with a high level of sustainable expertise in food composition. Sources of data used within the datasets include: analytical data produced specifically for the dataset using standard methods in accredited laboratories; data from scientific or grey literature, calculations from ingredients of composite foods, manufacturers' data and data from other national datasets. Data is managed and published using a range of data handling tools, most commonly relational databases, although some datasets are still compiled and published in spreadsheet format. Although data standards do exist, the structure of food composition data is not yet fully standardized and depends on the source of the data and the knowledge and expertise of data compilers.

There have been many collaborative projects and networks of food composition data compilers that have aimed to improve consistency and harmonization of composition databases, so that values from different datasets are of comparable quality. European projects such as EuroFOODS, Cost Action 99, the IARC European Nutrient Data Bank project (Slimani et al., 2007a; Slimani et al., 2007b) and the work of the INFOODS (International Network of Food Data Systems) organization (www.fao.org/infoods) network all made progress towards more standardized production,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

compilation and management of data. These and other related projects, summarized in 'The production, management and use of food composition data' (Greenfield and Southgate, 2003) were used as the basis for the European Food Information Resource (EuroFIR) project that aimed to standardize and harmonize food composition data in Europe through improved data quality, food description, database searchability and standards. EuroFIR produced a range of tools to help data compilers, including procedures for documenting data values, and supported the development and publication of a European standard for food data (EN 16104:2012, Food data - structure and interchange format) that was based on the EuroFIR technical standard (Becker et al., 2008; Becker, 2010; CEN, 2012). The CEN standard also took into account recommendations of the GS1 GDSN Trade Item Standard Food & Beverages extension (GS1, 2018) used in the retail industry, and the European Food Safety Authority (EFSA) Guidance on Standard Sample Description for Food and Feed (EFSA, 2010) that applies to chemical contaminants and residues included in monitoring and control programs. The CEN standard is therefore a flexible standard that can support not only food nutrients and bioactives data exchange, but also data on feed and data concerning other food properties (e.g. allergen or micro-organism contents, pH, vitamin retention factors).

The use of standards for data compilation and publication allows the development of a range of tools that help users to access and exchange data more easily and link to purchase, preparation and consumption data. These standards and tools should be used as the basis for use of food composition data within the RICHFIELDS Data Platform. Data capturing aspects of diet, health and lifestyle is generated through large scale research studies and also by individuals using web-based technology. Data structure is not standardized but many applications are collecting the same kind of data and there is overlap with composition, consumption and clinical data. Where web-based applications are used by consumers, the data collected will in most cases be in the format required by the app. The data may also be in response to feedback provided by the application at various stages of use.

The food composition standards developed for food classification, food description and description of component values are compatible with most systems for capturing and describing food intake data. Therefore, applications that provide data to the RICHFIELDS Data Platform should either directly structure data according to these standards or bespoke data structures used by applications should be mapped to these standards. Figure 2 illustrates the determinants of dietary behaviour, food consumption and nutrition links between different types of data related to food intake and nutrition, and highlights many of the data categories that could be included in the RICHFIELDS platform. Data can also be categorised by type, such as reference data (e.g. food composition data), observational data (e.g. food intake data, physical activity), or data that is transformed into output data (e.g. nutrient intake, dietary patterns).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

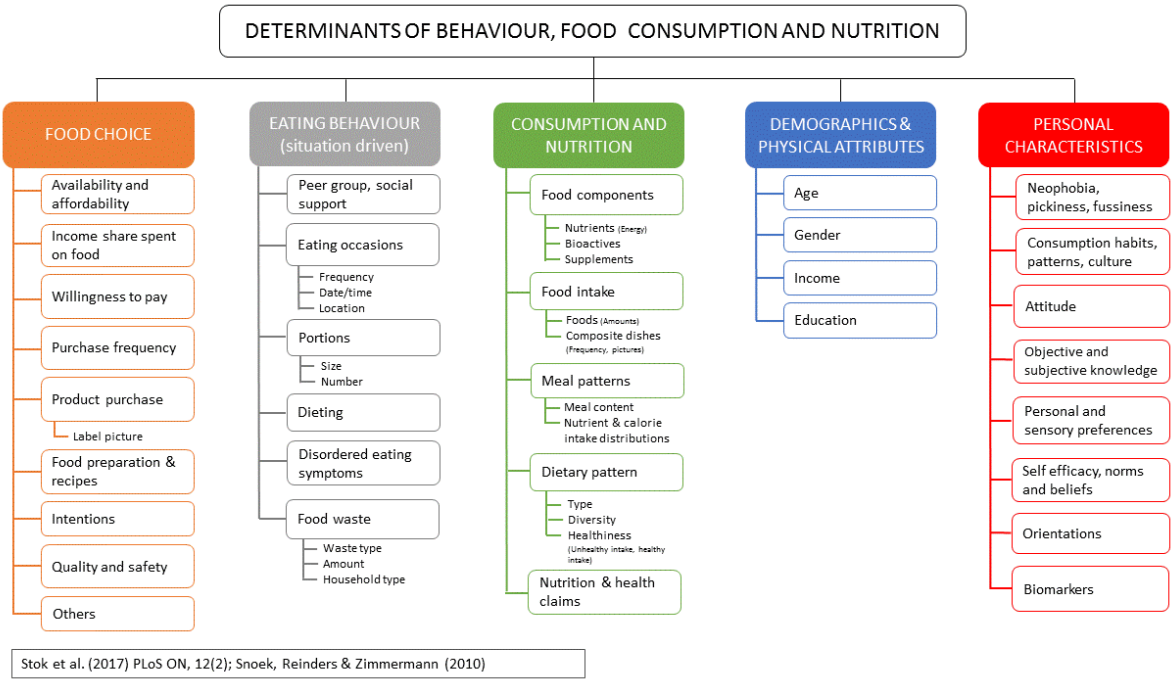

Figure 2. Overview of the determinants of dietary behaviour, food consumption and nutrition. The figure is adapted from the work by Stok et al (2017), undertaken as part of the DEDIPAC Knowledge Hub, and highlights the diversity of consumer data that may be collected or handled via the new RI.

5.2 Data storage and availability

EuroFIR originally envisaged a system where food composition data provided by the FoodExplorer tool was held locally by the data providers/owners and was accessed directly using web services, i.e. data was transferred directly between computer systems with no user intervention once the data request was made. However, while possible in a few cases, the proposed system relied on data providers having access to advanced IT infrastructures with sustainable IT expertise. In most cases producers of national food composition datasets could not provide the necessary resources and further intervention was needed. Even though food composition datasets are relatively harmonized and standardized, some reformatting is usually needed for further use and the current EuroFIR model stores modified data on a central server that allows access to data that can be presented in a more harmonized format. Despite the significant efforts to harmonize and standardize data, more work is needed and systems are continually evolving to improve data quality and ease of use. This is particularly important where food composition data is linked to other types of nutritional data such as food intake, contaminants, allergens or to biomarkers related to health. Continued provision of data and services has relied on a committed network of members and users. A sustainable IT and governance infrastructure and access to relevant expertise, including technical, administrative and management is essential.

The PRECIOUS platform implemented a data management and storage process to ensure reliable and secure management and storage of end user data collected by the user application. The platform allows sharing the data with external parties only if the user provided consent during the installation of the application. The PRECIOUS platform stores user data securely and keeps it protected and not

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

visible to external parties unless the user provides the consent for data sharing, which is indicated during the installation of the application. The PRECIOUS RI has gone through several ethical reviews and approvals carried out by the different partners including hospital and other national entities. Storage of data from user applications is however a very complicated issue and storage of and access to data provided via user applications to the RICHFIELDS Data Platform will need to be carefully considered.

The platform designed and developed by the PRECIOUS project partners, and evaluated in the case study, is composed of healthcare-related services that target physical activity, diet and motivation to follow a healthier lifestyle. The platform can be divided into three main components in which functionality is clearly differentiated. Firstly, end user data collection is performed by a smartphone application, environmental sensors or third-party sensors such as wearables or professional biomedical devices; this data is sent to the cloud using a dedicated server and stored into the user database which is called Virtual Individual Model Database (*VIM DB*). Secondly, user data is processed in the cloud by the *VIM* event trigger according to rules that are stored into the Rules Database; after processing, feedback is sent to the end user's smartphone. Lastly, experts (psychologists, physicians, diet specialists and others) design and create rules that under certain circumstances trigger events that can lead to different actions such as suggestions of healthier lifestyle or positive feedback in terms of smartphone notifications. Also, the *VIM* access server enabled field experts to access user data in order to detect early stage risks of non-communicable diseases due to monitored risk factors or create better and personal recommendations. This approach could be used to deliver, store and arrange access to data provided to the RICHFIELDS Data Platform.

5.3 Maintenance and access

Maintenance of data provided by and to the RICHFIELDS Data Platform could be a major task for the RICHFIELDS RI, depending on models of data delivery used. Providing and maintaining data that is up-to-date is a considerable challenge, particularly in the academic area where progress tends to be related to funding. Technology is relatively stable and cheap and therefore human resources are most likely to be the limiting factor.

The approach to maintenance of national food composition datasets means that a percentage of data will always be 'old' and some old data will not be representative of foods that are currently consumed. Ideally, data included in datasets will include information on the date of generation and/or publication and will be validated at the time of publication. Even where information on data validity is available, datasets will almost certainly still contain out of date values and validity of the dataset will decrease with time following publication. The impact of out of date values depends on what the data is used for but can be very significant in some cases. Datasets of branded foods may be updated continually as a result of agreements between producers and wholesalers and retailers, who often require updated composition data as a pre-requisite for distributing or stocking products. This information is often supplied by a third-party organization that receives information from producers and compiles the data into datasets that are then supplied to retailers for online use. An example is Brandbank who compile and distribute data, on a commercial basis, to provide online solutions for retailers. Although the values included in datasets are as up-to-date as possible, the format of the data and consistency

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

of food description poses some challenges to users. The data relate to very specific products, based on GTIN code, and new products are continually added while others are discontinued.

For most purposes it will be necessary to adapt published food composition data for specific uses. Most uses will require values for all nutrients to avoid the possibility of missing values being assumed to be insignificant and treated as zero. Some nutrients, e.g. fatty acids and individual sugars, are usually only provided for a limited sub-set of foods so may also need to be added by users. This is not a problem when users are only using data from one country but can be problematic when data from different countries and sources are combined. Many users of the data are not aware of the limitations of published data, despite detailed documentation being provided for most national datasets. Where users are aware of the need to adapt data, there is often a gap in the knowledge and skills and/or the resources needed to adapt the datasets for correct use. Branded food data usually only provides nutrient values that are required or permissible for food labelling. It is possible for branded foods to be mapped to generic foods so that mineral and vitamin contents can be estimated but that work requires considerable expertise and technical data management skills. For that reason, branded datasets have not been completed (in terms of nutrient coverage) and fully merged with generic national datasets.

The limitations of compiling datasets mean that there may also be limitations for users that depend on the purpose. 'Static' datasets that are included in applications may be out of date when they are used and in many cases data users do not routinely update to newer versions of the data. While updating is technically possible, many users do not have the skilled IT resources or sufficient knowledge of the data to make the necessary updates. For some purposes, e.g. research studies over extended time periods, it may not be desirable to use updated data because nutrient intake may then be changed because of changes in the underlying composition dataset rather than just in the foods consumed. An API approach to data linking has the advantage of being able to deliver up-to-date data that continually represents the data as currently published, compared with the traditional static datasets that may include out-of-date information. Access to branded food data, combined with generic food composition data, is likely to be an important feature of the RICHFIELDS Data Platform and would be a great benefit to researchers and commercial users.

For many applications, both research and commercial, dietary intake data needs to be linked to food composition data. One of the big challenges is to accurately match foods as consumed with data that accurately reflects their composition. Within applications the match is usually made by a simple food name search of the composition database and users make their own judgement on the composition data to use. Common problems with this approach include no obvious match being available and lack of detail to distinguish between food items that appear similar. Consistent food description and classification has always been a challenging problem in food composition and consumption databases. The LanguaL food description system (Moller & Ireland 2017) has been available for many years and can enable consistent language independent description and interpretation of foods in datasets that include LanguaL codes. More recently, EFSA has introduced and developed a comprehensive classification and descriptive system (FoodEx) which has been improved based on evaluations by users in a range of European countries (EFSA, 2015) and benefiting from the GloboDiet experience and its facet-descriptor concept. The use of FoodEx2 codes would solve some of the problems identified with

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

usability of food matching tools but further improvements would be needed when applied to more complex composite dishes using FoodEx2. In addition to developing improved tools for food matching, it is essential to develop the quality of data and information in both food composition datasets and food consumption databases.

In order to improve data linkage between GloboDiet (and other tools that record dietary intake) and the new RICHFIELDS Data Platform, it has been recognised that conceptual development of a more technical interfaced and partly computerised system for matching food consumption and food composition is needed. This would allow for greater alignment of procedures, cost-efficiency in resource needed for food matching, and would contribute to further standardisation of dietary methodologies and their related consumption data across Europe. Such an approach takes advantage of gathering new knowledge from human experts using the system, complex information from various food datasets and also new ICT knowledge.

The GloboDiet methodology for standardized food consumption data is fully compliant with EU menu recommendations for dietary software for dietary assessment (EFSA, 2009, Gavrieli et al., 2014) and it has already been customised, validated and tested for risk assessment and food safety, through different EU projects i.e. EFCOVAL, PANEU, PANCAKE. As part of the World Health Organization (a UN agency), IARC has developed general conditions to get access to and use the GloboDiet methodologies, endorsed by the GloboDiet dietary monitoring users. These general conditions aim to address three major issues: ensure the dissemination of international research tools and cumulated related knowledge while ensuring their standardisation over the time; acknowledge the contribution and data ownership of the GloboDiet users; acknowledge that the GloboDiet software and hard infrastructures are still under development and depend on broader European roadmaps. In this context, IARC provides the international umbrella to ensure standardisation, provides implementation support for dietary monitoring to the countries and ease the use and exchange of the international GloboDiet methodologies between users. IARC has made substantial investment in the GloboDiet research platform in the past to permit a wider use of GloboDiet in nutritional surveillance. However, because such a platform needs significant further investment and longer-term infrastructural support, and in light of IARC's available resources and competing research priorities, it was decided that IARC will not continue to invest in GloboDiet and that a new host for the platform would be sought. In June 2016 it was decided that the GloboDiet platform should be transferred to a new host with the capacity to develop/oversee nutritional surveillance programs at both the national and international level. The WHO Head Quarter (contact: Dr Francesco Branca, Geneva) has indicated strong interest in hosting the GloboDiet platform in the future. The technicalities of the setting-up of this platform are being currently evaluated.

One of the problems with sustainability of IT based tools is the cost of continuing to maintain and/or develop tools that are sustainable within constantly evolving IT platforms and this issue should be considered for development of the RI. Open source software and access to the original source codes are really important in order to maintain and improve software tools at affordable costs. The harmonization of food composition datasets included within GloboDiet does, however, demonstrate that different types of data can be linked between tools and platforms where there is a common harmonized data link. Such harmonization will be very important for successful data linking within the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

RICHFIELDS Data Platform. It is more likely that a range of dietary intake assessment tools would be a better approach for RICHFIELDS.

Access to clinical data is currently limited by a lack of standardization but the on-going works from ENPADASI WPs, particularly on structured and standardized data storage, actions required for clinical data sharing, could be the basis for all forecasted future RICHFIELDS development for clinical data integration. The ENPADASI project has specifically focused on the general rules to share and reuse data based on EU national policies, addressing ethics, data protection, data sharing policies and intellectual property. Within the ENPADASI project, the design of dedicated data sharing infrastructure was planned with related analysis of the data sharing issues. Combining local and central data sharing solutions facilitates long term sharing as it limits the resources needed and reduces duplications of data. RICHFIELDS will benefit from all this work that has been specifically dedicated to clinical data and bridges that could be built between RIs.

5.4 Ethical issues

The ethics of data collection and data transfer are an important issue for any data that relates to an individual, i.e. clinical or lifestyle (including food intake) data that has been collected as part of a research study or that is patient data. Ethical approval for research studies usually requires that the intended uses and mechanism of data storage and transfer are fully described in relation to each study, meaning that where data sharing with an RI is not specifically described in the ethical submission it may be difficult to gain retrospective approval.

It is likely that it will be more acceptable for individuals to choose to share their data for research purposes only rather than for commercial exploitation and the business model of the RICHFIELD RI may therefore have an impact on ethics. The proposed public-private model for the RI may cause some concern because of the possibility of data being available to businesses for commercial use and ethical use of data will need to be defined by policies for data access and sharing. If the public side of the RI has the dominant role in governance issues it would help to ensure close supervision of ethical issues and ensure the privacy and correct use of data supplied to RICHFIELDS.

ENPADASI together with CORBEL/ELIXIR has made a full overview of the ethical, privacy and IP issues related to data sharing. Evaluation of the processes used within the EUROSTARCH case study and RIs linked to clinical and health related data showed that interventional study data can be shared within an RI's platform infrastructure. A common ontology and quality standards are a minimum requirement for optimal and sustainable use of data so that data can be reused for new research questions, such as re-analysis, analysis of pooled data or secondary analysis. Anonymization of data is likely to be a pre-requisite and a need for significant reformatting is also highly likely, both of which require time and technical resources.

Review of the many relevant connections and similar issues among RIs, particularly ENPADASI and CORBEL, have demonstrated the complex regulatory context linked to clinical data use and re-use and the major importance of using proper informed consent. The optimal balance between optimizing data sharing and minimizing risks of jeopardizing privacy and researchers' legitimacy remains a key challenge. Ultimately, the broader access to data will strengthen transparency and standardization and help to raise researchers' awareness of the necessity to consider data sharing

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

aspects prior to project design and data collection. It will allow easier aggregation of data sets for secondary or meta-analysis or to be re-examined, to reach sufficient statistical power, improved generalizability, cross validation and to increase cost-efficiency of research data. Moreover, the European Commission as well as some journals have now put data sharing as a condition for funding and publication; this is also an opportunity to build on their novel scientific policies requiring open access to data.

A broad informed study participant consent makes it possible to use non-anonymous data for a broad range of scientific questions but ethical challenges regarding sharing of data must be considered in the mainstream of the RICHFIELDS platform design. Access to data is likely to be made easier when working within the framework of scientific networks that can help to facilitate sharing data and resources.

5.5 Recommendations for Research Infrastructure development

The EuroFIR model for harmonization, access to and use of food composition data has proved to be sustainable since the EuroFIR not-for-profit business was formed from the Network of Excellence project in 2009 and has been used as one of the business models in the RICHFIELDS project. The focus of the EuroFIR project and the EuroFIR AISBL has been providing harmonized data to users (research users and industry) as well as providing knowledge and training in the use of food composition and related data. EuroFIR links to other research networks and infrastructures through research and/or commercial projects. The expertise within EuroFIR allows it to not only provide food composition data but to provide expertise on the use and management of the data for a wide range of users and applications. The EuroFIR AISBL business model has been developed to provide expertise for both producers and users of food composition data, mainly in the research and industry (including food producers and software developers) sectors. While EuroFIR initially focused on providing harmonized food composition data to users, there has been an increasing need for EuroFIR to use data provided by other users and stakeholders. In particular, food composition data is often linked to food consumption or food purchase data that is produced either from large surveys (national monitoring or specific research activities) or generated by consumers and provided either through software applications or by industry (e.g. retailer) generated data. Figure 3 illustrates an example of how tools, data and components of the four case studies, related to food composition and food consumption, can contribute to a Research Infrastructure for Food, Nutrition and Health (FNH-RI) with links between platforms that provide data and/or tools (e.g. EuroFIR, MetroFoods) and RIs with indications of likely data flows between them. There is a wider landscape that can include many other networks, data sources and tools that could also be part of the FNHRI.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

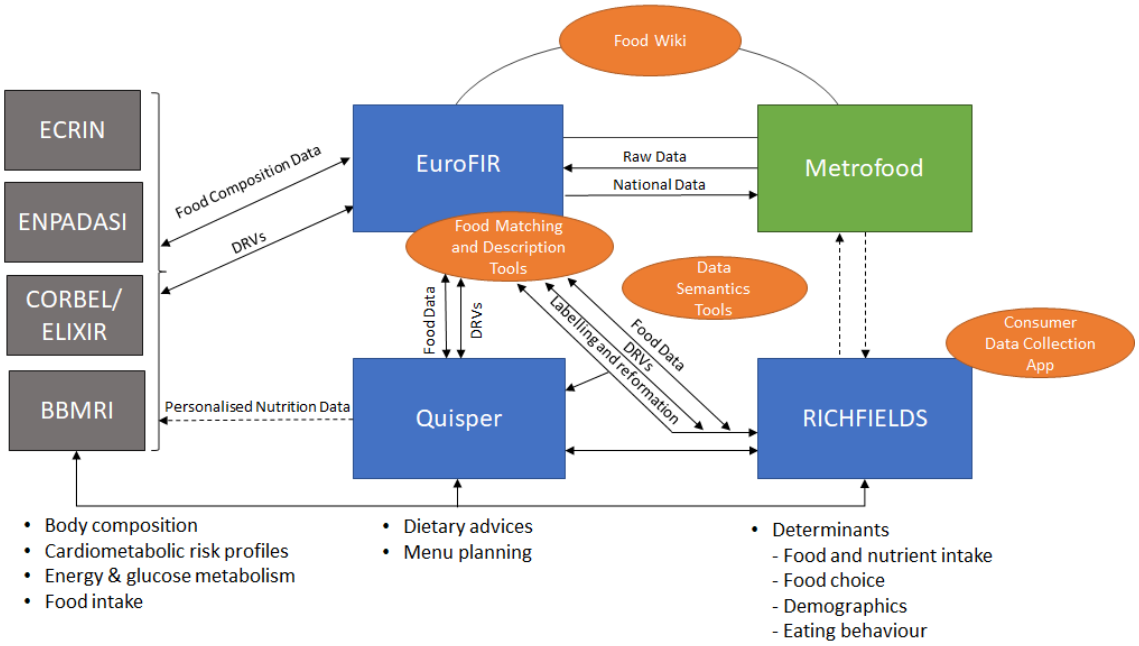

Figure 3. Example of how existing and emerging RIs and RI-like structures related to the case-studies (food composition and food consumption, determinants and personalized dietary advice) could be linked and share data in the future.

Since the provision of data to users will be central to RICHFIELDS, the EuroFIR experience described in this case study provides an example of how data and services can be provided by a network that relies on data being made available by external data providers. EuroFIR uses a modular approach to providing services that has evolved over a few years. The model caters for use by a range of users including both researchers and industry users. Sustainability is achieved through a mixed model of membership fees and 'on demand' fees for some licensed data and consultancies (including participation in research projects). Fees are structured so that larger organizations and businesses pay more compared to SMEs or individual users.

EuroFIR relies on use and re-use of data that is provided by external organisations. In many cases, the data is made freely available as a result of governmental or organizational policies. However, there are some data that are the subject of other agreements between data providers and EuroFIR. These agreements may change over time as policies dictated by the data providers change and it has been important for EuroFIR to maintain relationships with these data providers to continue to provide access to a wide range of data. EuroFIR has tried to add value to the data, e.g. by harmonizing data, improving quality, or providing access to networks or tools to improve data management. Assuming that data used by and provided to RICHFIELDS will not initially be generated by the RICHFIELDS Data Platform itself, it will be essential for RICHFIELDS to make agreements with data providers to allow use. In cases where data is available under an open licence, agreements should be straight forward but in other cases bespoke agreements will be necessary, perhaps with associated costs for access to the data. EuroFIR has found that where data providers have a commercial interest (actual or perceived) in their data, it is unlikely that data will be provided without an agreement reflecting that value.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

The case studies reported examples of networks/RIs that have been set up based on different models. These models have had different levels of success to date and served as examples to inform development of RICHFIELDS. The broad range of data providers and users envisaged to participate in the RICHFIELDS platform will provide challenges and, based on the EuroFIR and QuaLiFY (The Qualify Project, 2018) experiences, focus on specific user groups (e.g. researchers, software providers) is likely to help enable development of simpler and more sustainable data access, governance and business models. The QuaLIFY project, which included EuroFIR AISBL as a partner, also investigated potential business models to support a proposed QuaLIFY business consortium. The concept was to develop and manage a unique gateway through which end-users could acquire access to a set of customizable services that deliver data and knowledge to support the development of personalized dietary advice to consumers. A key concept was the identification of data ownership and possible restrictions on data transfer as an important issue and it was agreed that the responsibility for physical storage, including maintenance of data and the gateway APIs would remain with data providers. Even this approach has not yet produced a sustainable entity following completion of the project because of difficulties associated with establishing suitable governance arrangements.

The RICHFIELDS Data Platform should focus on the collection, harmonization and linkage, analysis and presentation of data to users. The system will enable access to data from one or more sources and knowledge extraction from the data, however the original data is not intended to be modified. Current links between national food consumption and food composition datasets (in many countries compiled by the same organizations) and users of the data are shown in Figure 7. Governance and administration aspects of the system will be crucial to allow systems managers to manage the RICHFIELDS data and knowledge repository while the system itself should be managed by system administrators. The RICHFIELDS Data Platform is proposed to be a stand-alone system that is connected with other information systems and e-infrastructures (e.g. EuroFIR information platform, Global Data Synchronisation Network (GS1 (GDSN)), ELIXIR, ENPADASI, etc.) via a range of web services.

Table 3 outlines the strengths and weaknesses of the data types that have been identified and evaluated by the four case studies. The strengths and weaknesses listed suggest a range of opportunities that could be available to RICHFIELDS. These are outlined in Table 4. Some of the opportunities listed in Table 3 are already included in some existing RIs and have been incorporated in successful and sustainable business models. To meet the needs of a wide range of research and commercial users and uses, it will be necessary for RICHFIELDS to link to existing relevant and sustainable RIs and to develop a range of flexible business models. Modular approaches would allow a combination of models to be used including membership based, subscription based, pay as you go services and consultancy projects.

The system should aim to support the governance model (i.e., management/administration and business model/monetization), while some other functionalities (such as big data storage and processing, semantic modelling, analytical services, etc.) may be provided from outside the system as well as some within the system. Platform users should include all the system's stakeholders, its managers, and other information systems. The system’s stakeholders should either provide data to the system or use data provided by the system while the managers should process data and maintain

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

and support the system with data being accessed by the information systems, including web services. It is essential that the system operates according to relevant organizational policies and external regulatory requirements that will affect the operation and performance of the system. The simplest approach appears to be to make virtual links to other RIs and link to the data that is held elsewhere. This would avoid some of the issues related to data transfer and ownership but access to data and business arrangements would still need to be addressed.

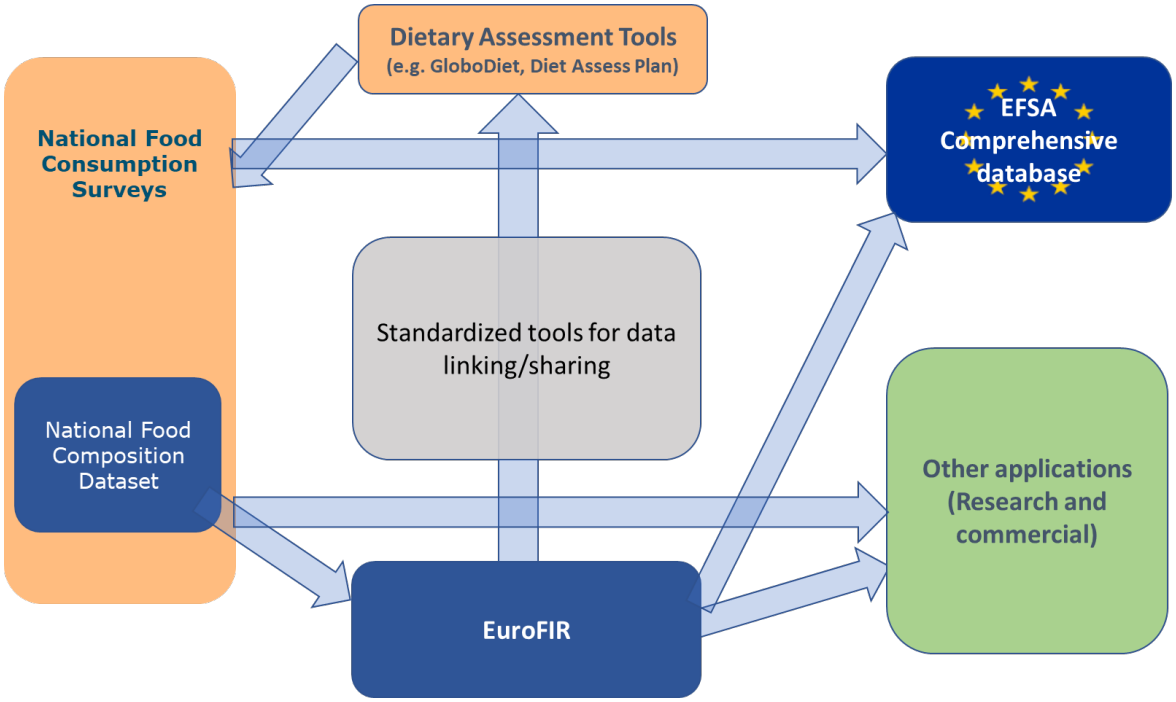

Figure 7. Example of data from individual datasets being provided to data platforms and then being made available to larger RIs and through to a range of end users and stakeholders.

6 Conclusions

The four case studies reported were evaluated to address the key issues related to design of the RICHFIELDS Data Platform, and the following conclusions, related to the key evaluation questions, can be made:

6.1 Data structure:

- Common ontology and quality standards are a minimum requirement for optimal and sustainable use of data
- Food matching systems and tools are vital to link datasets that are based on food names and descriptions, e.g. food composition and food consumption datasets. Food classification and description systems, e.g. FoodEX2, LanguaL, need to be included in datasets to allow the food entities to be matched
- Automatic approaches to food matching have been developed and are promising but food matching cannot yet be completely automated and further development is needed
- Training data providers and users will be essential to enable effective use of shared data

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

- Use of clinical or lifestyle data originating from patients or individuals will need to conform to ethical standards.
- Transfer of clinical or lifestyle data from a data provider to RICHFIELDS will require anonymization
- Even where data provided to the platform is harmonized there may still be a need for significant data reformatting

6.2 IC-technology and data storage:

- Dataset formats should not be an important issue because any database (or spreadsheet) system used for food and lifestyle data should be compatible with easy data manipulation and import/export to other formats, e.g. xls, mdb, sql, xml, cvs
- Anonymization and harmonization of data both require time, technical and scientific investments
- Costs of maintaining access to data need to be considered
- The review of ENPADASI, particularly on structured and standardized data storage and requirements for clinical data sharing, serves as a good example for using clinical data within a RI
- Combining local and central data sharing solutions could facilitate long term data because it maximizes use of resources and limits duplication
- The PRECIOUS project, used as a case study for lifestyle data, demonstrated that web-based technology can be used to generate data and transfer data to other users.
- The principle of inclusion of data from consumers or businesses does not yet seem to have a clear example that could be utilized directly by RICHFIELDS to engage and incentivize individuals and organizations to make their data available.
- The selected case studies have demonstrated that structures are already in place to allow linking between existing RIs and the RICHFIELDS Data Platform.

6.3 Governance and ethical issues:

- National and international policies, privacy, practical and regulatory issues are necessarily linked with clinical data collection and use, and use of human data is particularly sensitive
- In many cases, to allow data sharing it will be essential for data providers to describe the governance of data and any possible follow up use (e.g. by an RI) as part of the research ethics application for the project
- The optimal balance between optimizing data sharing and minimizing ethical and legal risks will be a key challenge
- Data access policies and control procedures will be needed to approve access to and use of data
- Data ownership with respect to business use, IP and data outputs (e.g. citation policies) will be important and policies should be produced to allow ease of use and promote sustainability
- Development of the platform as either a public RI, a private RI or a public-private RI will have significant implications for platform management, data provision and access and will have an impact on scientific, social and economic impacts of the RI

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

- Governance arrangements will have an impact on viable business services that can be included in the core value proposition
- The operation and governance of RIs and organizations that provide data to RICHFIELDS may dictate the way that RICHFIELDS can provide data

References

Ainsworth BE, Haskell WL, Leon AS, Jacobs DR, Montoye HJ, Sallis JF, et al. Compendium of Physical Activities - Classification of Energy Costs of Human Physical Activities. Med Sci Sport Exer. 1993 Jan;25(1):71-80. PubMed PMID: WOS:A1993KG53700011. English.

BBMRI-ERIC. BBMRI-ERIC Policy for Access to and Sharing of Biological Samples and Data 2018 [22/08/18]. Available from: http://www.bbmri-eric.eu/wp-content/uploads/AoM_10_8_Access-Policy_FINAL.pdf.

Becker W, Data CTF. Towards a CEN Standard on food data. Eur J Clin Nutr. 2010 Nov;64:S49-S52. PubMed PMID: WOS:000283752600010. English.

BioMedBridges. Legal and Ethical Requirements Assessment Tool - LAT 2018 [22/08/18]. Available from: http://www.biomedbridges.eu/supporting-researchers-sharing-sensitive-data-identifying-requirements.

Brown KA, Timotijevic L, Geurts M, Arentoft JL, Dhonukshe-Rutten RAM, Fezeu L, et al. Concepts and procedures for mapping food and health research infrastructure: New insights from the EuroDISH project. Trends Food Sci Tech. 2017 May;63:113-31. PubMed PMID: WOS:000400532800011. English.

Bucher T, van der Horst K, Siegrist M. The fake food buffet - a new method in nutrition behaviour research. Brit J Nutr. 2012 May 28;107(10):1553-60. PubMed PMID: WOS:000304222900020. English.

dbNP Project. dbNP: The Phenotype Data Infrastructure 2018 [22/08/18]. Available from: http://www.dbnp.org/.

de Boer EJ, Slimani N, van 't Veer P, Boeing H, Feinberg M, Leclercq C, et al. The European Food Consumption Validation Project: conclusions and recommendations. Eur J Clin Nutr. 2011 Jul;65:S102-S7. PubMed PMID: WOS:000292448100013. English.

Eftimov T, Korosec P, Seljak BK. StandFood: Standardization of Foods Using a Semi-Automatic System for Classifying and Describing Foods According to FoodEx2. Nutrients. 2017 Jun;9(6). PubMed PMID: WOS:000404177100013. English.

Eftimov T, Korousic Seljak B. POS Tagging-probability Weighted Method for Matching the Internet Recipe Ingredients with Food Composition Data. In Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR, (IC3K 2015)2015.

Eftimov T, Koroušić Seljak B, Korošec P, editors. Grammar and Dictionary based Named-entity Linking for Knowledge Extraction of Evidence-based Dietary Recommendations. IC3K 2016 Proceedings of the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management; 2016.

Eftimov T, Seljak BK, Korosec P. A rule-based named-entity recognition method for knowledge extraction of evidence based dietary recommendations. Plos One. 2017 Jun 23;12(6). PubMed PMID: WOS:000404145100017. English.

EFTIMOV T, ISPIROVA G, FINGLAS PM, KOROŠEC P, KOROUŠIĆ-SELJAK B. Quisper ontology learning from personalized dietary web services. In: AVEIRO, David (Ed), DIETZ, Jan L. G. (Ed), FILIPE, Joaquim (Ed). *Proceedings. Volume 2, KEOD*. [S. l.]: SCITEPRESS = Science and Technology Publications. 2018, pp. 279-286.

European Commission. 2018. About Research Infrastructures [24.09.2018]. Available from: https://ec.europa.eu/research/infrastructures/?pg=about

European Food Information Resource AISBL. Welcome to EuroFIR AISBL 2018 [22/08/2018]. Available from: http://www.eurofir.org/.

European Food Safety Authority. General principles for the collection of national food consumption data in the view of a pan-European dietary survey. EFSA Journal. 2009;7(12).

European Food Safety Authority. Standard sample description for food and feed. EFSA Journal. 2010;8(1):1457.

European Food Safety Authority. The food classification and description system FoodEx2 (revision 2). EFSA supporting publication 2015 [Internet]. 2015; EN-804:[90 p.]. Available from: https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/sp.efsa.2015.EN-804.

European Nutritional Phenotype Assessment and Data Sharing Initiative. European Nutritional Phenotype Assessment and Data Sharing Initiative 2018 [22/08/18]. Available from: http://www.enpadasi.eu/.

EUROSTARCH. What is EUROSTARCH? 2018 [22/08/2018]. Available from: http://www.eurostarch.org/.

Finglas PM, Berry R, Astley S. Assessing and Improving the Quality of Food Composition Databases for Nutrition and Health Applications in Europe: The Contribution of EuroFIR. Adv Nutr. 2014 Sep;5(5):608s-14s. PubMed PMID: WOS:000342972200026. English.

French Obesity Research Centre of Excellence. Accueil 2018 [22/08/18]. Available from: https://www.force-obesity.org/.

Gavrieli A, Naska A, Berry R, Roe M, Harvey L, Finglas P, et al. Dietary monitoring tools for risk assessment. EFSA supporting publication 2014 [Internet]. 2014; EN-607:[287 p.]. Available from: https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/sp.efsa.2014.EN-607.

Greenfield H, Southgate DAT. Food Composition Data: Production, Management and Use. Second ed. Rome: Food and Agriculture Organization of the United Nations; 2003.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

GS1. Welcome to GS1: The Global Language of Business 2018 [22/08/18]. Available from: https://www.gs1.org/.

International Network of Food Data Systems. About INFOODS 2018 [22/08/2018]. Available from: http://www.fao.org/infoods/infoods/en/.

Jimoh F, Lund EK, Harvey LJ, Frost C, Lay WJ, Roe MA, et al. Comparing Diet and Exercise Monitoring Using Smartphone App and Paper Diary: A Two-Phase Intervention Study. Jmir Mhealth Uhealth. 2018 Jan;6(1). PubMed PMID: WOS:000433951300013. English.

Mezgec S, Seljak BK. NutriNet: A Deep Learning Food and Drink Image Recognition System for Dietary Assessment. Nutrients. 2017 Jul;9(7). PubMed PMID: WOS:000406679700019. English.

Mezgec S, Eftimov T, Bucher T, Korousic Seljak B. Mixed deep learning and natural language processing method for fake-food image recognition and standardization to help automated dietary assessment. Pub Health Nutr. 2018. doi:10.1017/S1368980018000708. E-pub ahead of print.

Møller A, Ireland J. LanguaL™ 2017 – The LanguaL™ Thesaurus. Technical Report.: Danish Food Informatics; 2018. Available from: http://www.langual.org/langual_literature_LanguaLReports.asp.

Mutafungwa E, et al. D4.1 System architecture and design specification. The Precious Project, 2015. Available from: http://www.thepreciousproject.eu/wp-content/uploads/2013/12/D4.1-System-architecture-and-design-specification_Final.pdf

PAPA, Gregor, KOROUŠIĆ-SELJAK, Barbara, PAVLIN, Marko. Device and method for acquisition and transfer of signals: patent GB 2525403 A. London: Intellectual Property Office, 28th October, 2015.

Papa G, Koroušić Seljak B, Korošec P, Piletić M, Hren I, Pavlin M (2018). Innovative pocket-size Bluetooth kitchen scale. Agro Food Industry Hi Tech, TKS Publisher, in press.

Park MK, Freisling H, Huseinovic E, Winkvist A, Huybrechts I, Crispim SP, et al. Comparison of meal patterns across five European countries using standardized 24-h recall (GloboDiet) data from the EFCOVAL project. Eur J Nutr. 2018 Apr;57(3):1045-57. PubMed PMID: WOS:000427967500016. English.

Plumb J, Pigat S, Bompola F, Cushen M, Pinchen H, Norby E, et al. eBASIS (Bioactive Substances in Food Information Systems) and Bioactive Intakes: Major Updates of the Bioactive Compound Composition and Beneficial Bioeffects Database and the Development of a Probabilistic Model to Assess Intakes in Europe. Nutrients. 2017 Apr;9(4). PubMed PMID: WOS:000401355600005. English.

Slimani N, Casagrande C, Nicolas G, Freisling H, Huybrechts I, Ocke MC, et al. The standardized computerized 24-h dietary recall method EPIC-Soft adapted for pan-European dietary monitoring. Eur J Clin Nutr. 2011 Jul;65:S5-S15. PubMed PMID: WOS:000292448100002. English.

Slimani N, Deharveng G, Unwin I, Southgate DAT, Vignat J, Skeie G, et al. The EPIC nutrient database project (ENDB): a first attempt to standardize nutrient databases across the 10 European countries

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

participating in the EPIC study. Eur J Clin Nutr. 2007 Sep;61(9):1037-56. PubMed PMID: WOS:000249276900001. English.

Slimani N, Deharveng G, Unwin I, Vignat J, Skeie G, Salvini S, et al. Standardisation of an European end-user nutrient database for nutritional epidemiology: what can we learn from the EPIC Nutrient Database (ENDB) project? Trends Food Sci Tech. 2007;18(8):407-19. PubMed PMID: WOS:000248632900002. English.

Slimani N, Ferrari P, Ocke M, Welch A, Boeing H, van Liere M, et al. Standardization of the 24-hour diet recall calibration method used in the European Prospective Investigation into Cancer and Nutrition (EPIC): general concepts and preliminary results. Eur J Clin Nutr. 2000 Dec;54(12):900-17. PubMed PMID: WOS:000165965600008. English.

Slimani N, Valsta L, Grp E. Perspectives of using the EPIC-SOFT programme in the context of pan-European nutritional monitoring surveys: methodological and practical implications. Eur J Clin Nutr. 2002 May;56:S63-S74. PubMed PMID: WOS:000175556100010. English.

The European Committee for Standardisation (CEN). European Standard. Food data - structure and interchange format. EN16104:2012 2012 [cited EN16140:2012]. Available from: http://www.sis.se/en.

The PRECIOUS Project. Welcome to PRECIOUS 2018 [22/08/18]. Available from: http://www.thepreciousproject.eu/.

The PRECIOUS Project. Precious Trial App. Google Play Store. Version 3.002, 2017. Available from: https://play.google.com/store/apps/details?id=aalto.comnet.thepreciousproject&hl=en.

The Qualify Project. Qualify – Quantify Life Feed Yourself 2018 [24/09/2018]. Available from: http://www.qualify-fp7.eu/

The RICHFIELDS Project. Welcome to RICHFIELDS 2018 [22/08/18]. Available from: https://www.richfields.eu/.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No. [654280]

Table 1. The RICHFIELDS linkage and harmonisation techniques and tools

Problem	Technique/tool	Reference	Partner
Matching foods classified using different systems	A new semi-automatic system, named StandFood, that aims to standardize foods according to FoodEx2 classification and description.	Eftimov et al, 2017	JSI
Identifying portion size	A pocket size Bluetooth weighing scale, named Libra.		JSI/QIB
Estimating nutritional impact of food choice using the Fake Food Buffet	A tool to measure food product and portion size choice in laboratory studies. A technique for estimating the fake food composition.	Mezgec et al, 2018	ETHZ, JSI
Linking GS1 data for branded foods to generic food composition data	A method known as drNER, which automatically recognizes food, nutrient, and quantity/unit concepts from unstructured text.	Eftimov et al, 2016; Eftimov et al, 2017a	JSI
Recognising food consumed from images	NutriNet, a Deep Learning food and drink image recognition system.	Mezgec and Koroušić Seljak, 2017	JSI
Searching for food composition data across multiple databases	FoodEXplorer, an interface that allows simultaneous searching for information across most EU food composition datasets, plus the US and Canada. Uses harmonised descriptions to link foods and nutrients (LanguaL).	Westenbrink et al (2018); Finglas et al (2014)	EuroFIR, CAPNUTRA, Quadram
Exporting composition data from multiple sources	Food Data Transport Package, a webservice utilising a standardised XML template that aides the import and export of food composition data.	Westenbrink et al (2018); Finglas et al (2014)	EuroFIR, JSI

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

Table 2. The focus and data outputs of RIs and networks related to data standardization and sharing

RI	Focus/data collection and outputs
ECRIN	Observational, interventional, and clinical data used Anonymized and/or codified data /informed consent Sampling as a key factor for data quality Increased use to open access External access to be approved by expert panel Patients given access to data via server + defined policy for access Need for repositories of cohort studies
ECRIN Nutrition (A project of ECRIN)	Nutrition hub since 2012 – 80 clinical centres in 21 countries. Possible areas for collaboration with FNHRI/RICHFIELDS include cooking kitchens, dietary tools (via meduniwien), online FFQ builder. Dietary tools for patients and cohort studies, multinational & randomized studies. Can provide insights and best practice for instruments/services, respective/prospective studies, data format, storing and access, ethics and user access requirements.
ENPADASI	Interventional and observational data Follows FAIR (Findable, Accessible, Interoperable and Re-usable) principle for data sharing Has federated databases across Europe Interactive analysis Developed nutrition ontologies Other specificities have been described throughout the deliverable 9.2 document.
CORBEL/	Network of 11 RIs in biological and medical sciences led by ECRIN/BBMRI/ELIXIR Sharing data, developing integrated joint services and stakeholder engagements Infrastructure service Integration of Platform Shared Services (Health & Biosciences) - use cases and community engagement (Data Management and Integration, Innovation, ELISI, Physical Access and Integration) Training Project management & communication

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

Table 3. Outline of the strengths and weaknesses of the data types that have been identified and evaluated by the four case studies

Data type	Strengths	Weaknesses
Food composition	• Authoritative and comprehensive national composition datasets freely available to users • Integrated into EuroFIR RI • Standards and thesauri available for food composition data • Data for branded products easily available • Bioactive compound datasets available • Some relevant RIs already exist (EuroFIR, INFOODS)	• Composition datasets may not be suitable for all uses ○ Foods and/or nutrients may be missing ○ Some data may be out of date ○ Software tools/apps may not use best available data ○ Software tools/apps may use data incorrectly ○ Users may be unaware of data limitations • Branded data usually available only for labelling nutrients • Quality of branded food data not clear • Bioactive compound data not linked to nutrient composition data • Available standards/ontologies not always used
Intake	• Aggregated national consumption datasets available • Large food purchase datasets collected by commercial organizations • Relevant RIs already exist • Software tools enable collection of end user data (diet/health/lifestyle) • Standardized tool (GloboDiet) for dietary intake assessment used for national monitoring in some European countries ○ Similar tools found to produce comparable results in EFSA study • Other software used across Europe (EU Menu)	• Accurate food matching tools needed to link intake and composition data • Portion size estimation significant limitation of intake data • 'Raw' consumption datasets not easily available • Purchase information not easily available to researchers • Not clear what current RIs can offer and sustainability not clear • End user data from software tools not usually available for research • Available standards/ontologies not always used
Clinical/biological	• Clinical/biological data (anonymised) published in scientific journals • Relevant RIs already exist • Data can be collected from electronic devices/sensors	• Clinical/biological data not available in compiled or standardized datasets • No easy access to data • Lack of standards/resources to interpret and make clinical decisions on user generated data

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

		• Not clear what current RIs can offer and sustainability not clear • End user data from software tools not usually available for research • Available standards/ontologies not always used
Health and lifestyle	• Relevant RIs already exist • Software tools enable collection of end user data (diet/health/lifestyle) • Data generated by wearable technology can be accessed and transferred by smartphone technology	• Not clear what current RIs can offer and sustainability not clear • End user data from software tools not usually available for research • Standard/ontologies not developed or available

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.

Table 4. Potential opportunities that could be available to RICHFIELDS

Standards
• Maintain standards for compilation and exchange of composition and intake data for research • Develop/harmonize standards for health/lifestyle data • Promote use of standards and thesauri in food and nutrition research • Offer training in production and use of standards/ontologies
Data sources and resources

• Expertise to improve use of data and accessibility to wider datasets (consumers and App/software developers)
• Develop improved tools/apps for users
  ○ Portion size estimation
• Improve access to and quality of composition data for branded products
  ○ Calculate content of additional nutrients
  ○ Develop/utilize APIs to access/exchange data
• Link bioactive compound data to nutrient composition data
• Improve access to consumption datasets produced by research/national monitoring
• Improve access to food purchase datasets for research use
• Improve access to standardized/compiled clinical/biological data
• Enable researcher access to end user generated data

Training

• Offer training in
  ○ Use of standards
  ○ Use of data
  ○ Production and exchange of data
• Targeted training for different user groups
• Dietary advice to consumers and other users for more personalized nutrition services

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654280.